A Proposal to Investigate the Effect of Felon Disenfranchisement on Recidivism

Erin M. Ochoa

MACS 30000: Perspectives on Computational Analysis

2016 October 24

Research question

Labeling theory posits that the mere act of labeling someone as a criminal can foster future criminal behavior. For previously convicted persons, the stigma of a felony conviction and societal barriers to reintegration may have a similar effect. Even after incarceration and supervision end, ex-felons retain the label. Depending on their state of residence, they must contend with many or all of the collateral consequences that accompany felon status. Among these is losing the right to vote.

We consider that, as a barrier to reintegration and a continuous reminder of the label, felony disenfranchisement could shape ex-felons' actions and perhaps foment criminal behavior. Given that crime is universally considered a social problem, it is in the interest of societies to avoid implementing policies or laws that increase crime. Therefore, we wish to investigate the following question: What is the effect of felon disenfranchisement on recidivism?

Research plan

State-level differences in felon disenfranchisement laws offer a prime opportunity to carry out a matching study. For the purposes of this study, we will consider felon disenfranchisement as the treatment condition. The expected outcome of treatment is recidivism.

With regard to felony-related voting restrictions, states fall into one of five[1] general categories: 1) No such restrictions; 2) inmates barred from voting; 3) inmates and parolees barred from voting; 4) inmates, parolees, and probationers barred from voting; and 5) inmates, parolees, probationers, and ex-felons barred from voting. These indicate control (no disenfranchisement; these are the states in category 1) and treatment (felon disenfranchisement; these are the states in categories 2 through 5). Among states in the treatment groups, there are four dosage levels.

There are a variety of ways to measure whether an ex-felon recidivates within a given period of time; these include, but are not limited to: Technical probation or parole violation, rearrest, reconviction, and reincarceration. Another strategy is to use survival analysis to measure the effect of treatment on the time elapsed between prison release and recidivism (defined as one of the aforementioned events).

Methods

For the purposes of this study, we will define recidivating as either of the following: reincarceration due to revocation of probation/parole or a new felony conviction. Because of the different ways to conceptualize both the treatment condition and the outcome, we propose using a mixed approach, assessing the effect of dosage level on both whether recidivism occurs and in time to recidivism. In both approaches, the independent variable, control versus four dosage levels of treatment, will be operationalized nominally. This is for two reasons: 1) Conceptually, each level of voting restrictions is distinct from the others and affects progressively larger proportions of the population as the severity grows; therefore, these dosage levels should be considered separately, not together; and 2) effect sizes are likely to be small, but we expect to see a bigger effect from the most severe dosage than from less-severe dosages; if we were to average the effects of different dosages together, we might not find a significant effect when one does, in fact, exist. In the first approach, the dependent variable, whether a given subject recidivates, will be operationalized dichotomously; in the second approach, the dependent variable, time elapsed between release from prison and recidivating, will be operationalized continuously in days.

We will employ a two-level matching design. First, we will construct a sample of at least two states from each category, as identified above; because there are only two states in category 1, they

---

[1] Christopher Uggen, Sarah Shannon, & Jeff Manza. "State-Level Estimates of Felon Disenfranchisement in the United States, 2010." The Sentencing Project. 2012 July. Washington, D.C. http://www.sentencingproject.org/wp-content/uploads/2016/01/State-Level-Estimates-of-Felon-Disenfranchisement-in-the-United-States-2010.pdf

will both be included. The sample will comprise at least ten states. We will attempt to match states on the following characteristics: Population density, proportion urban, percent Black, median income, average educational attainment, percent incarcerated or under supervision, and proportion registered to vote. We recognize that this strategy is ambitious and may require revision once implementation begins. Second, we will match individual offenders from each state to individual offenders in the matched states. We will attempt to match offenders on the following characteristics: Age, gender, race, ethnicity, and most serious convicted offense. There may be a surplus of individuals who cannot be matched; these will be omitted.

There are important reasons to use a two-level matching design. First, using a matching design at all is preferable over not as it may reduce the influence of unknown or unobserved variables. Second, a single-level matching design—one that matched only states—could lead to an ecological fallacy in which we would be making inferences about individual behavior based on state-level characteristics. In this study, the unit of analysis is the individual, so we cannot be satisfied with simply matching states.

Data and sources

Because of the state-by-state nature of felon disenfranchisement laws, we will need data from states in each category, as identified above, in order to detect differences between treatment and control conditions as well as detecting differences between dosage levels.

For each state, we will need the following data sets: a) Prison-release records dated 2004, b) probation-termination records dated 2004, c) parole-completion records dated 2004, d) voter registration records dated 2005–2015, and e) conviction records dated 2005–2015.

We will limit the scope of our inquiry to persons released from or under the supervision of state corrections systems, excluding persons solely under the jurisdiction of the federal corrections and probation systems.

Record linkage

We will need to link together at least five data sets per state. We know from experience that this is difficult but not impossible. We will develop a probabilistic linking framework employing phonetic algorithms, machine learning, name rarity[2] scores, and typo approximation[3] to uniquely identify offenders across multiple data sets.

Justification

This study exemplifies the use of digital data and computational methods. Specifically, it leverages the favorable characteristics of big data and effectively compensates for its limitations.

Favorable characteristics of big data

Big: As mentioned earlier, the effect of felon disenfranchisement on recidivism is expected to be small. This is precisely why we must use big data to approach this question. Only by combining multiple large data sets from various agencies in multiple states can we hope to uncover a small yet potentially significant effect. The data sets to be collected for this study would together comprise a very long database of many individual offenders.

Always on: Because the data sets, by their very nature, capture every change in status as offenders move through the criminal justice system (or, for voter registration data sets, every

---

2 This will be a method to weight names inversely to their frequency, either in a data set or in a corpus of names (possibly ranked by year). Because we thought of it just the other day during a conversation about tf-idf, we have not yet defined an algorithm. We are not aware of any similar methods currently in use for names.

3 This is an untested method we began developing earlier this year to match individuals across two or more data sets. It involves intentionally introducing typos into birthdates and social security numbers to attempt to match individuals despite typos. We are hopeful that it will allow us to match individuals who would otherwise go unmatched. We are not aware of any similar methods currently in use.

instance of voters registering), they are always on. Though there is a lag between an event occurring and the creation of the corresponding digital record, the data sets capture every eligible event that happens over time. The data can therefore be used for longitudinal analyses, as in this study. In this sense, the data sets are always on.

Non-reactive: The data required are non-reactive. Having one's personal identifiers collected is a normal, if bureaucratic, part being arrested, jailed, imprisoned, released, or paroled. We know of no suggestion that the mere act of collecting such information influences behavior.

Limitations of big data

Incomplete: Alone, each required data set is incomplete. It is only by bringing the data sets together that we can harness the power of government-collected big data. Each data set is created by a different state agency and collects only the information the agency needs to fulfill its purpose. However, failure to obtain even a single necessary data set could affect the execution of the study.

Inaccessible: Because each data set is owned by a different state agency in at least ten states (at least eight of which have not yet been determined), the data will not be easily accessible. By partnering with other agencies, however, we can increase the likelihood that we will obtain the data. We will use existing ties with the New Mexico Sentencing Commission and the New Mexico Statistical Analysis Center to make connections with sentencing commissions and statistical analysis centers in other states. Additionally, we will network with former professors and academic mentors who may have professional relationships with others who may have access to the data sets or know someone who does. Finally, we will attempt to form a partnership with an academic entity, such as a laboratory or research institute, and leverage its professional connections to obtain the data.

Non-representative: The main concern about the data being non-representative is related to neither incompleteness nor random sampling but rather to a methodological choice: discarding unmatched offenders. With an appropriately small proportion, this would be only marginally concerning, but a large enough proportion of unmatched offenders would raise questions about the effectiveness of the state-level matching. While it is beyond the scope of this paper to determine the specific matching methodology to be employed, an alternate technique might be to attempt matching of offenders between states by random sampling until a sample of a given size has been generated.

Drifting: Because the data will come from so many different agencies, we will almost certainly have to contend with differently structured data sets containing intersecting sets of variables; consider this a type of interagency drift. We have dealt with such data sets before and know that these challenges are not insurmountable.

Dirty: As discussed above in the context of record linkage, the data will be dirty, but not in the way that Salganik[4] defines "dirty".

Sensitive: The data needed for this study are sensitive. While conviction records are public, personal identifiers such as offenders' birthdates and social security numbers are not. In order to gain approval from an institutional review board, we must have a plan to securely obtain, store, and eventually destroy the data. Because the data are digital, we must have access to a secure computer server to store the data; to gain access to such a machine, we must partner with an academic institution or faculty member. This is possible and even likely.

Feasibility

As outlined herein, we believe that this study is feasible, though certainly complex. Given the implications the results could hold for law, public policy, reintegration, and recidivism, we believe it is needed and important.

[4] Matthew J. Salganik, *Bit By Bit: Social Research in the Digital Age* (http://www.bitbybitbook.com, in review), 2.3.2.6.